研究主题视域下零被引与高被引论文分析*

——以环境科学领域为例

■ 潘菲 王效岳 白如江 周彦廷
山东理工大学科技信息研究所 淄博 255049

摘要：[目的/意义]从主题视角对环境科学领域的零被引论文进行分析，对比零被引论文与高被引论文在文章内容、外在指标方面的不同，揭示零被引论文存在的原因。[方法/过程]首先，对来自 Web of Science 数据库的国内环境科学领域的 260 篇高被引论文、907 篇零被引论文的摘要进行 PLDA 主题识别，然后通过主题相似度计算发现主题间的关联，以主题热度作为内部指标，发文时间、发文期刊作为外部评价指标，最后，把论文主题内容与外部指标结合进行零被引与高被引论文之间的相同主题、不同主题对比分析。[结果/结论]在研究主题相同情况下，期刊的影响因子大小是影响零被引论文的主要因素；在主题不同的情况下，论文研究的主题内容是导致零被引论文的主要原因。

关键词：零被引 高被引 主题识别 环境科学 对比分析 评价指标
分类号：G250
DOI：10.13266/j.issn.0252-3116.2018.20.009

1 前言

随着科学技术和信息技术的发展，科技文献数量不断增加，文献引用频次的研究受到研究人员的高度关注。在文献计量学中，衡量论文影响力或质量的基本指标是论文的被引次数[1-2]。相对于高被引论文而言，零被引论文的关注度低，研究热度远远不及高被引论文，但零被引或低被引文献的潜在价值，如果被发掘出来，也许远远大于我们目前的想象[3]。零被引论文是指一个国家、机构、学科、期刊或个人在某年或某个时间周期内出版的论文集合，在出版后的某一个或几个不同长短的引用时间窗口中未受到任何引用的论文[4]。关于为何存在零被引论文以及如何更好地发现零被引论文的价值，是学者们关注的问题。本文针对高被引与零被引论文提出以下几个问题：高被引论文因研究高热度主题而引用频次高，还是因发表在影响因子大的期刊上而被引频次高？零被引论文因主题研究热度低而被引次数为零，还是因发表在影响因子小的期刊上而无被引？针对这些问题，本文选择环境科学领域的高被引论文和零被引论文数据，利用 PLDA (Parallel Linear Discriminant Analysis)模型，即并行隐含狄利克雷分布模型进行主题识别，结合论文的发文时间、发表期刊等指标进行分析，评价各指标对零被引与高被引论文的影响，以发现影响零被引论文的因素及其规律。

2 相关研究

20 世纪 50 年代美国著名情报学家 E. Garfield 提出了引文分析方法，也是情报分析和科学评价的常用方法。但是以往的各类指标都主要倾向于关注引用分布曲线上代表“高被引论文和受关注论文”的头部，却没有关注“低被引论文和暂时无人关注的论文”。根据长尾理论，零被引论文对科学界的贡献与高被引论文所做的贡献相匹配。2004 年英国学者 A. Weale 等[5]提出可将零被引率(non-cited rate)作为期刊质量反向评价指标，随后 T. N. Van Leeuwen 和 H. F.

* 本文系国家社会科学基金项目“未来新兴科学研究前沿识别研究”(项目编号:16BTQ083)和山东省软科学重点研究计划项目“深化高校、科研院所科研体制改革对策研究”(项目编号:2017RZB01046)研究成果之一。
作者简介：潘菲(ORCID:0000-0003-1574-7353)，硕士研究生；王效岳(ORCID:0000-0002-7100-7758)教授，博士，硕士生导师，通讯作者，E-mail:wangxy@sdut.edu.cn；白如江(ORCID:0000-0003-3822-8484)，副教授，博士，硕士生导师；周彦廷(ORCID:0000-0001-7624-2637)，硕士研究生。
收稿日期:2018-03-15 修回日期:2018-06-08 本文起止页码:77-87 本文责任编辑:刘远颖

Moed 发现期刊影响因子与期刊论文零被引率之间存在下降的函数关系，两者的皮尔逊相关系数为负 0.63[6]。国内学者唐晓莉以经济学科为例验证了零被引率用于期刊反向评价是合理的[7]。李美玉等认为图书情报领域验证零被引率可以作为期刊关键评价指标的反向指标，但要考虑学科差异[8]。通过研究零被引率与期刊评价指标之间的相互关系，并基于此构建新的、融合零被引率的科研评价指标，扩展了期刊评价的标准，肯定了零被引论文的价值。为了对零被引论文进行全面研究，学者们对零被引论文产生的原因和特征进行了分析。

关于学术论文得不到引用，除了论文本身学术水平外，是由很多因素造成的。方红玲以我国 5 种眼科学中文核心期刊 2003 年发表的零被引论文为研究对象，分析发现在下载量和被引量关系中，部分低被引甚至零被引论文具有较高的下载量，在主题分布中，零被引论文主题分布广泛[9]。魏瑞斌等认为造成零被引的主要原因是论文选题太偏，不属于主流研究领域[10]。高继平等以 JCR 光谱学期刊为例，认为数据统计来源、论文发表时间、研究主题等是零被引论文的重要影响因素[4]，杨思洛以图书情报档案学科 15 种核心期刊为例，认为零被引论文率均值与篇均被引率、H 指数负相关，与综合排名正相关；零被引论文率在不同时间、期刊、学科间差异明显[11]。胡泽文采用问卷调查法寻找零被引的原因，结果表明论文发表时间短、论文质量不太高、论文主题偏冷门或不够新颖、所发期刊的影响力（或质量）较低是出现零被引的主要原因[12]，温芳芳以情报学期刊的论文为研究对象，通过零被引与高被引论文的比较，认为论文的可见度和可获得性、作者的影响力、论文合著者数量以及论文选题是否新颖和热门等因素，均对论文被引频次产生不同程度的影响[13]。杜新征等[14]从论文类型、内容结构、基金分布、作者机构分布、页码和语种 6 个方面分析《水生生物学报》零被引论文的特征，发现页码和语种与零被引没有相关性，其他指标都可以反映零被引论文特征。赵越从主题因素分析，发现零被引论文的研究主题分散、陈旧，但是研究主题并没有显示出偏离学科研究领域或者高度前瞻性的特点[15]。况书梅等以图书情报领域的论文关键词为主对论文零被引进行分析，在被引论文与未被引论文方面，二者的研究相似度逐渐下降，且研究内容与学科热点的相关程度很大程度上影响零被引的出现几率[16]。李贺琼等对 10 种外科学综合类期刊 2011 年零被引论文进行分析，发现零被引率与影响因子关系不大，署名 2－5 位作者占大多数，第一作者所属机构以省市级医院为主，零被引论文无基金资助占大多数，论文类型以临床研究和病例报告为主[17]。

总体看，研究学者从数据本身对零被引论文出现原因进行相关研究，包括论文的类型、国家和机构科研实力、基金分布、学科差异、语种、科研合作程度、文章选题等因素，引文计数虽是研究论文的重要指标，但仅是简单的数据统计分析，并没有深入到数据背后对应的论文主题信息。即使从主题因素进行研究，研究学者仅停留在关键词、高频词等文本信息的研究，没有深入到论文内容中，缺乏语义之间的联系。本文选择环境科学领域的高被引论文和零被引论文数据，利用 PLDA 模型进行主题识别，以主题热度作为内部指标评价，结合论文的发文时间、发表期刊等外部指标进行分析，研究分析零被引论文存在的原因和规律。

3 研究思路

为了进一步分析零被引论文与高被引论文之间的区别与联系，笔者将论文的研究主题作为内部评价指标，将发文时间、发文期刊等作为论文的外在评价指标，只有对数据进行由内而外的分析，才能有效分析零被引论文的特征，而主题模型是分析论文内部信息的有效手段。本文的研究思路分为 3 个步骤，如图 1 所示：

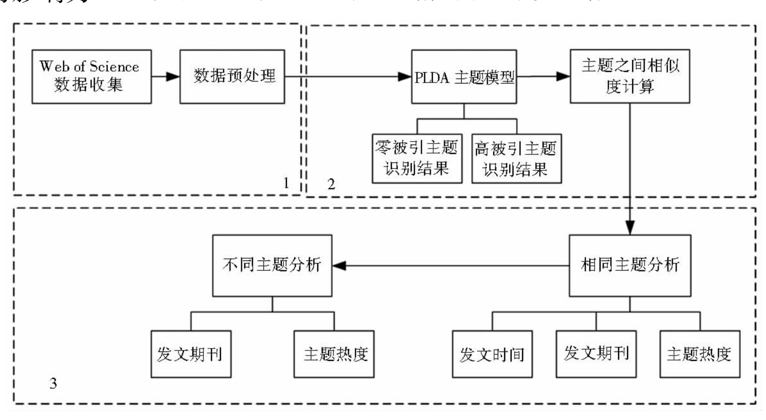


图 1 高被引与零被引论文对比分析研究思路

（1）数据收集与预处理。首先确定数据的来源，选择 Web of Science 数据库进行数据的搜集，根据构建的检索式获取该学科领域的文献。然后对获取的文献进行预处理和数据的清洗，包括去除停用词、标点符号及数字，提取词干等步骤，同时对该领域的高频词和文献中无意义的词语进行处理，留下有研究意义的词语，

为之后的文本主题识别提供支持。

（2）主题识别与相似度计算。利用 PLDA 模型识别出蕴含在高被引论文摘要和零被引论文摘要中的主题，并构建主题－文档和主题－主题词矩阵，根据主题－主题词矩阵，利用主题相似度计算方法对零被引论文的主题与高被引论文的主题进行相似度计算，发现零被引论文主题与高被引论文主题的区别与联系。

（3）特征提取与分析。PLDA 主题模型的识别结果提供了主题下的论文发文时间、发文期刊等数据，根据提供的主题及主题下的数据，运用主题热度、发文时间、发文期刊等指标得出影响论文零被引的主要因素，为下一步分析零被引论文的原因提供方法与思路。

3.1 PLDA 主题模型

PLDA 模型是基于 Gibbs sampling 近似分布并行框架的 LDA 模型，为保证主题数量的准确性，选择统计语言模型中常用的评价指标即困惑度（perplexity）确定主题的最佳数量，困惑度越小，主题识别越好。D. M. Blei 等定义了一个有 M 篇文档的文档集的主题模型的困惑度[18]为：

$$perplexity(D_{test}) = \exp\left\{ -\frac{\sum_{d=1}^{M} \log p(w_d)}{\sum_{d=1}^{M} \log N_d} \right\} \quad 公式(1)$$

其中，M 是文档集中的文档的数目，$P(W_d)$ 是 LDA 模型生成第 d 篇文档的概率，N_d 是第 d 篇文档的单词的个数，当 perplexity(D_{test}) 最小时，K 个主题最能够表达模型的语义关系，即可以确定 No of topic 主题数。

3.2 主题相似度

通过 PLDA 主题模型识别出来的主题，主题内部之间的联系可以用主题相似度进行分析。本文用余弦相似度计算主题之间的相似性并设置阈值 Y，相似度大于 Y 则认为两个主题之间相同，否则视为不同。第一步构建向量空间模型（Vector Space Model，VSM），把高被引主题与零被引的主题用向量的方式进行描述，向量空间模型中用 T（Topic）表示主题、T（Term）表示主题词、W（weight）表示主题词权重，主题向量可用主题词表示为 Topic = ｛t1，t2，t3 …… tn｝、主题词权重向量为 Topic Vector = ｛w1，w2，w3 …… wn｝，每个主题词都有一个权重；第二步计算两两主题之间的相似度，计算结果介于[0，1]之间，数值越大相似度越高。主题相似度计算公式为：

$$Sim(Topic_i, Topic_j) = \cos\theta = \frac{\sum_{k=1}^{n} w_k(Topic_i) \times w_k(Topic_j)}{\sqrt{(\sum_{k=1}^{n} w_k^2(Topic_i)) \times (\sum_{k=1}^{n} w_k^2(Topic_j))}} \quad 公式(2)$$

其中，分子表示两个主题向量的点乘积，分母表示两个主题向量模的乘积。

3.3 零被引与高被引论文对比分析指标

本文借鉴目前研究中提出的混合式判断指标[19-20]，通过分析、总结论文的文本内容、外在属性等特征，提出主题热度、发文时间、发文期刊的论文分析指标体系，通过指标体系的构建分析零被引论文现象。

3.3.1 主题热度指标　论文的发文量、被引量可以作为研究热度的评价指标[19-20]，根据本文识别出的每个主题下论文数量，将主题热度定义为每个主题内部论文数量，即通过统计不同主题内部论文数量占总论文数量的权重，以表征各个主题的热度，主题热度能够直观地分析研究主题的关注度、影响力变化趋势[21]。

计算公式为：

$$TH = \frac{X_i}{\sum_{j=1}^{n} X_j} \quad 公式(3)$$

其中，TH 代表主题热度（topic heat）；X_i 代表每个主题下的论文数量；$\sum_{j=1}^{n} X_j$ 代表所有主题下论文数量之和。

3.3.2 发文时间指标　发文时间指标，是分析指标中的基础因素，主要分析主题下不同年份论文数量变化，根据论文数量下的年份反映主题发展趋势，是新生、成长还是消亡[22]，从而看出研究学者对论文的关注度。

3.3.3 发文期刊指标　科技期刊是论文的主要载体，期刊质量的高低也影响论文的被引次数，本文通过零被引论文与高被引论文的期刊对比，得出期刊对论文的影响。本文运用 SPSS 的指数回归进行分析。指数模型的计算公式为：

$$Y = \beta_0 e^{\beta 1x} \quad 公式(4)$$

公式(4)中，Y 为每个主题下每个期刊的载文量，x 为按照时间排列论文顺序，$\beta_0\beta_1$ 为常数。

4 实验

4.1 数据源与预处理

4.1.1 数据源　Web of Science 具有权威性、完整性等多种优势，因而本文从 Web of Science 数据库中进行数据采集，并选择环境科学领域作为研究学科。检索数据库：SCI-EXPANDED；数据检索式：TI = “environ＊”；时间跨度：2006 －2015 年；文献类型：article and review；检索类别及研究方向：environmental science & environmental sciences ecology；检索国家：Peoples R China；检索语种：English；检索时间：2017 年 6 月 20 日；检

索结果:15 002 篇。对检索结果进行初步分析,根据零被引论文的定义,本文将 2006 – 2015 年 10 年间被引频次为零的论文作为零被引论文;高被引论文则以汤森路透集团[23]文献评价分析工具 ESI 为主,将高被引论文(most cited papers)定义为过去 10 年被引用次数排在各学科前 1% 的论文。最后得到零被引的论文 907 篇,高被引论文数量为 260 篇。

通过图 2 发现环境科学领域高被引论文和零被引论文数量都表现出增长趋势,高被引论文在 2006 年为 0,10 年后,高被引论文数量达到 58 篇,说明在此期间环境科学专业有较好的发展;零被引论文在 2006 年数量为 19 篇,到 2015 年论文数量达到 464 篇,是 2006 年论文数量的 24 倍,增长幅度大于高被引论文的增长幅度。分析其原因主要有:①论文通常会在发表的一至两年后达到引用高峰,因此零被引的数量较多;②随着论文数量的增多,不相关的学者生产出与之不相关的论文,很难查阅到所有相关论文,导致论文不被引用。



图 2　高被引论文与零被引论文数量变化折线

4.1.2　数据预处理　利用文献题录信息统计分析工具 SATI3.2 对高被引和零被引论文摘要进行提取,然后利用 Python 对论文的摘要部分进行文本数据的预处理和清洗,包括标点符号和数字剔除、大小写转换、停用词处理、词干提取等,之后再次对文本数据进行清洗,去掉与主题不相关和该领域的高频词,如 environment、environmental、china、Elsevier、right、paper 等,提高主题识别的准确度。结果见图 3。

```
chitin[]abund[]natur[]occur[]polysaccharid[]dispos[]
seafood[]crustacean[]mainli[]shrimp[]prawn[]lobster[]
compon[]shellsexoskeleton[]crustacea[]widespread[]chemi[]
physic[]versatil[]materi[]value-added[]applic[]
chitosan[]initi[]develop[]wide[]rang[]deriv[]highli[]
stabl[]difficult[]degrad[]obtain[]shell[]suitabl[]
process[]abl[]potenti[]treatment[]remov[]wastewat[]deacetl
[]produc[]structur[]research[]look[]adsorb[]industri[]leachat[]
identifi[]deffici[]exist[]equilibrium[]
carri[]determin[]capac[]variou[]kinet[]solut[]methodologi[]adopt[]
explain[]adsorpt[]knowledg[]requir[]design[]commerci[]system[]
pressur[]varieti[]direct[]caus[]automobil[]suppli[]manag[]consid[]initi[]
implement[]practic[]improv[]econom[]perform[]expand[]earlier[]explor[]pressuresdriv[]
motiv[]automot[]empir[]enterpris[]experienc[]regulatori[]market[]
strong[]intern[]driver[]adopt[]especi[]consider[]extern[]relationship[]slightli[]
oper[]
specif[]
organ[]
```

图 3　论文摘要文本处理结果

4.2　零被引论文主题与高被引论文主题对比分析

4.2.1　实验参数设置与结果分析　对文本数据进行处理之后,要对数据进行主题识别,主题识别的准确性与主题数量有很大关系,重要的参数设置为主题数量(No. of topic)和主题下的主题词数量(No. of words per topic)。本文对主题数量 No. of topic 和困惑度 perplexity 对应关系进行实验。由于高被引论文摘要文本量较少,预估主题数量在 15 个以内,主题数量 No. of topic 设为 2 – 12 个,按步进量为 2 进行处理,得到主题数和困惑度对应关系,如图 4 所示。从图 4 中可以看出,当主题数量 No. of topic 为 10 时折线变化趋势逐渐稳定,虽然主题越少困惑度越小,但会造成过度拟合,因此最终确定主题数量 No. of topic 为 10,每个主题选择 15 个主题词。其他相关参数设置:Alpha0.5;Beta0.1;迭代次数 2 000。

同理,对零被引论文的摘要也进行了相应的实验,主题数量为 12 时,主题困惑度趋于稳定。

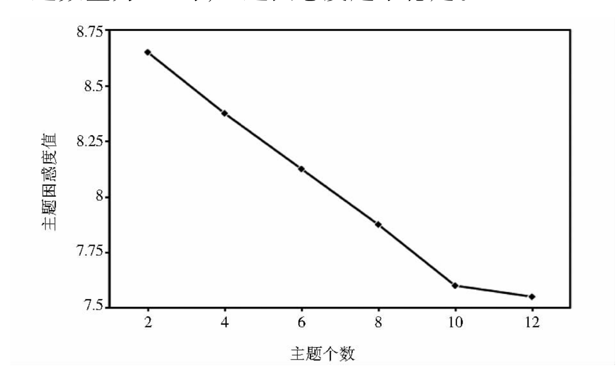


图 4　高被引论文主题数和主题困惑度关系

零被引与高被引论文主题识别结果如表 1、表 2 所示,不同的主题代表不同的研究内容,对识别的主题结果进行总结。零被引论文研究主题分 3 个类型:①环境污染物的研究。主题 0 是通过在水中暴露的浓度,研究污染物对水生生物和植物的毒性研究。②自然生态系统的评价研究。通过数学建模、物理建模、空间建模、景观生态模型法和各种监测机制,从整体的角度去评估生态系统的现状及变化,分析生态系统发生的变化及原因以及环境变化对环境自身和人类的影响,为制定合理的区域生态环境管理政策,遏制区域生态环境恶化,改善区域生态环境质量提供依据。主题 1、主题 2、主题 3、主题 4、主题 6、主题 11 是关于生态系统评价的研究,涉及气候、地质、水文等各方面的评价,其中主题 11 基于景观的评价体系是生态系统的新视角。③针对不同的污染物采用不同的治理方法。通过对主

题5、主题7、主题8、主题9、主题10的分析,发现该主题是对环境污染的原因进行分析并采用微生物－生物技术治理大气、河流、土壤中的污染物。

表1 零被引论文主题识别结果

主题	主题词
Topic0	Concentr, studi, level, effect, exposur, active, toxic, significantly, indic, decreas, methane, control, factor, speci, potenti
Topic1	chang, studi, region, climat, river, model, watersh, surface, process, eros, factor, season, sedim, distribut
Topic2	chang, wetland, studi, product, factor, effect, region, ecosystem, valu, season, forest, develop, indic, natur
Topic3	river, pollut, qual, method, assess, studi, model, indic, index, region, factor, concentr, base, health, ecolog
Topic4	develop, system, model, studi, manag, sustain, region, evalu, urban, chang, polici, paper, effect
Topic5	model, studi, growth, us, concentr, strain, investig, level, method, effect, health, radionuclide, degrade, wast
Topic6	model, predict, method, particl, reservoir, studi, surface, effect, urban, base, algorithm, concentr, simul, propos
Topic7	emiss, product, carbon, treatment, pollut, effect, studi, industry, energy, reduc, plant, mushroom, method, potenti, concentr
Topic8	bacteri, express, studi, effect, concentr, protein, activ, cell, level, metabol, detect, indic, exposur, sequenc
Topic9	sedim, metal, heavi, organ, concentr, soil, plant, microbe, studi, sampl, indic, content, carbon, nitrogen
Topic10	remov, adsorpt, concentr, process, solute, studi, oxid, effect, treatment, efficien, wastewat, method, condit, investig
Topic11	speci, spatial, studi, commun, ecology, plant, index, indic, region, divers, pattern, factor, protect

表2 高被引论文主题识别结果

主题	主题词
Topic0	adsorpt, surface, biochar, sorption, adsorb, magnet, magnetization, carbon, interact, graphen, complexes, investing, isotherm, effect, nanotub
Topic1	antibiotics, concentr, detect, resist, respect, treatment, tetracyclines, investig sulfonamide, wastewater, bacteria, street, effluent, Correlations
Topic2	material, effect, barrier, industrial, pollut, concentrations, contamin, strategi, increase, agricultural, soil, efficien, swidden, automot, option
Topic3	effect, assess, chemic, pollut, biochar, nanoparticles, monitor, develop, provid, mixtur, exposur, sorption, zebrafish, impact, contamin
Topic4	structur, effect, climat, degrade, signific, electron, measure, applications, impact, hierarchical, challenge, efficien, ecosystem, region
Topic5	carbon, develop, responsive, product, system, increase, community, reserve, capture, effect, micropollutants, global, photocatalyt, energy, mercury
Topic6	e-waste, pollut, recycl, forest, understand, review, develop, impact, ecosystem, potentiall, reserv, region, concentr, improve, chemic
Topic7	energy, efficiency, consumption, industry, performace, reserve, product, economic, technology, develop, growth, construct, measure, pollut, process
Topic8	degradation, pollutants, process, product, health, effect, concentr, increas, treatment, antibiotics, particles, contamin, identify, potenti
Topic9	photocatalyt, catalyst, graphen, applic, materal, energy, perform, electron, composit, structur, reaction, g-c3n4, exhibit, degrad

高被引论文主题主要分两类:①环境污染物本身的研究。如主题1抗生素污染问题、主题3各种纳米材料对环境的潜在危险研究、主题4各种污染材料的有效利用、主题6电子垃圾和重金属污染的处理,这些主题从不同的方面对环境污染物进行了研究;②对环境污染物治理的研究,主要从治理环境的材料和技术两个方面进行分析。环境材料对污染物的治理有主题0、主题2、主题7,主题0、主题2两个主题从石墨烯、碳纳米管等新型材料的吸附功能进行污染物治理研究,主题7从生态能源、绿色能源的角度以减少污染物的产生;对治理污染物进行技术研究的主题有主题5、主题8、主题9,这3个主题从生物治理方法、降解技术、光催化处理技术对各类污染物治理进行研究。

从主题识别的结果来看,零被引论文的主题除了在生态系统评价方面的研究,还包含高被引论文的研究主题,即污染物研究和污染物处理技术的研究,由此看来零被引论文的研究主题更加广泛。高被引论文的主题集中,研究采取具体的技术措施,强化对有毒有害的危险污染物的认识和治理方案。

4.2.2 主题热度 主题热度是对该领域的研究方向的反映,不同的主题会有不同的主题热度,主题热度的大小则代表了对主题的关注度,研究热度越高则关注度高,热度越低则关注度也低。设置主题热度的阈值为0.1,大于0.1则主题热度高,关注度高。在小于0.1的主题中,根据主题热度的大小,排名最后的3个主题为低热度主题,其他剩余主题为一般研究主题。

零被引与高被引论文的主题热度如表3所示:零被引论文中主题4、主题9、主题10为高热度主题,一般主题是主题0、主题1、主题3、主题6、主题8和主题11,主题2、主题5、主题7为低热度主题。同理,高被

引论文中主题 0、主题 1、主题 7、主题 9 为高热度主题，主题 3、主题 6、主题 8 为一般主题，主题 2、主题 4、主题 5 为低热度主题。

表 3　零被引与高被引论文主题主题热度计算结果

零被引主题	论文数量	主题热度	高被引主题	论文数量	主题热度
0	70	0.077 7	**0**	**37**	***0.141 8***
1	68	0.075 5	**1**	**28**	***0.107 3***
2	60	0.066 6	2	20	0.076 6
3	67	0.074 4	3	21	0.080 5
4	**101**	***0.112 1***	4	17	0.065 1
5	47	0.052 2	5	17	0.065 1
6	78	0.086 6	6	23	0.088 1
7	60	0.066 6	**7**	**45**	***0.172 4***
8	62	0.068 8	8	24	0.092 0
9	**94**	***0.104 3***	**9**	**29**	***0.111 1***
10	**133**	***0.147 6***			
11	61	0.067 7			

4.2.3　主题相似性　利用 Python 的 Gensim 工具包对零被引与高被引论文识别出的 22 个主题进行相似度计算，通过设置阈值，判定零被引与高被引论文主题的相似程度，主题相似度的计算结果见表 4。

通过相似度计算发现在高被引论文主题中，主题 9 跟零被引论文主题的相似度最少，HT9 是对各种材料作为光催化剂的研究，重点研究纳米材料，如石墨型碳氮化合物（g-C3N4），零被引论文主题中主要集中在微生物处理技术和生态系统评价中；而零被引论文主题中，主题 11 跟高被引论文的主题相似度最少，ZT11 是生态系统评价的研究，包括城市、农业、森林、海洋、景观生态等各个方面，该方法可以对各种潜在的土地生态影响进行评估，在高被引论文主题中没有涉及生态系统评价。为了将零被引论文与高被引论文放在相对平等的条件下进行对比，具有研究意义，将它们之间相似度的阈值设为 0.2，大于 0.2 即主题之间相同，小于 0.2 则不具有关系。通过表 4 得出相似的主题有 4 组，分别是 ZT2 与 HT4 关于各种污染材料的有效利用的研究，ZT4 与 HT5 是水污染控制与环境微生物技术研究，ZT7 与 HT8 两个主题研究内容为污染物处理技术，最后，ZT10 与 HT0 是关于环境吸附内容的研究，主要是吸附的技术和材料研究。剩余的其它主题则为不相似主题。

表 4　高被引论文主题与零被引论文主题相似度计算结果

	HT0	**HT1**	**HT2**	**HT3**	**HT4**	**HT5**	**HT6**	**HT7**	**HT8**	**HT9**
ZT0	0.062 5	0.064 5	0.062 5	0.125 0	0.064 5	0.062 5	0.066 8	0.000 0	0.187 5	0.000 0
ZT1	0.000 0	0.000 0	0.000 0	0.000 0	0.133 3	0.000 0	0.069 0	0.066 7	0.064 5	0.000 0
ZT2	0.129 1	0.000 0	0.064 5	0.129 1	*0.276 0*	0.193 6	0.200 0	0.133 3	0.129 1	0.000 0
ZT3	0.000 0	0.064 5	0.062 5	0.125 0	0.064 5	0.000 0	0.193 6	0.064 5	0.125 0	0.000 0
ZT4	0.066 8	0.000 0	0.066 8	0.133 6	0.138 0	*0.200 4*	0.142 9	0.069 0	0.066 8	0.000 0
ZT5	0.129 1	0.133 3	0.064 5	0.064 5	0.066 7	0.064 5	0.069 0	0.066 7	0.193 6	0.064 5
ZT6	0.064 5	0.066 7	0.064 5	0.064 5	0.066 7	0.064 5	0.069 0	0.000 0	0.129 1	0.000 0
ZT7	0.125 0	0.129 1	0.125 0	0.125 0	0.064 5	0.187 5	0.133 6	0.129 1	*0.312 5*	0.062 5
ZT8	0.064 5	0.133 3	0.064 5	0.129 1	0.066 7	0.064 5	0.069 0	0.000 0	0.129 1	0.000 0
ZT9	0.064 5	0.066 7	0.064 5	0.000 0	0.000 0	0.064 5	0.069 0	0.000 0	0.064 5	0.000 0
ZT10	*0.258 2*	0.200 0	0.064 5	0.064 5	0.133 3	0.064 5	0.069 0	0.066 7	0.200 0	0.000 0
ZT11	0.000 0	0.000 0	0.000 0	0.000 0	0.069 0	0.000 0	0.071 4	0.000 0	0.000 0	0.000 0

注：ZTO 为零被引主题 0；HT1 为高被引主题 1

4.3　零被引与高被引论文相同主题特征对比分析

4.3.1　零被引与高被引论文研究主题、发文时间对比分析　通过相似度计算，高被引与零被引论文有 4 组主题相同，对相同主题的零被引与高被引论文的研究主题、发文时间和发文期刊进行研究，主题的发文时间见表 5。

（1）HT0 与 ZT10 特征分析。

结论一：主题热度高。论证：从主题内容上分析，是关于对污染物吸附内容的研究，两个主题的主题热度高，共同关注该领域的研究热点。

结论二：发文时间不同。论证：HT0 的研究主题近 10 年都有引用，发文时间最多的一年是 2014 年；ZT10 在 2006 – 2011 年间零被引论文数量较少，2011 年之后才开始增长，分析其原因为：环境吸附问题作为高被引主题中关注度高的主题，引起了研究学者关注，产生了众多研究成果，导致一些论文还没有被引次数。

（2）HT5 与 ZT4 特征分析。

结论一：主题热度不同。论证：两个主题研究水污

表 5　高被引与零被引论文相同主题发文时间对比

主题	年份	数量	主题	年份	数量	主题	年份	数量	主题	年份	数量
HT0	2014	9	HT5	2014	4	HT4	2015	3	HT8	2014	7
	2015	6		2009	3		2010	4		2015	5
	2009	5		2010	3		2011	2		2011	4
	2010	5		2012	3		2012	2		2013	4
	2012	4		2013	2		2013	2		2009	2
	2008	3		2015	2		2014	2		2010	1
	2011	3					2007	1		2012	1
	2007	1					2009	1			
	2013	1									
ZT10	2015	68	ZT4	2015	42	ZT2	2015	27	ZT7	2015	37
	2014	23		2014	21		2014	14		2014	8
	2013	14		2012	12		2013	7		2013	5
	2012	11		2013	9		2012	5		2012	5
	2011	6		2011	6		2011	3		2009	2
	2009	5		2006	4		2010	1		2010	1
	2006	3					2009	1		2008	1
							2008	1		2007	1
							2007	1			

染控制与环境微生物技术，HT5 虽然出现在高被引论文中，但研究热度低，而 ZT4 为热点主题，关注度高。

结论二：论文发文时间靠前。论证：该主题的发文时间都集中在近 5 年，说明该主题作为短期的研究前沿，具有新颖性和先进性，但还没有形成体系，从而产生大量零被引论文。

(3) HT4 与 ZT2 特征分析。

结论一：主题热度低。论证：虽然两个主题共同关注各种污染材料的有效利用，但研究热度低。分析其原因为：污染物的处理技术研究难度升高，如厌氧技术、碳纳米技术等，需要相应的设备和条件才能推进发展。

结论二：发文时间具有可持续性。论证：该主题的发文在近 10 年都有被引用论文和未被引用的论文，主题研究难度高，突破性技术少，导致论文发表时间周期长。

(4) HT8 与 ZT7 特征分析。

结论一：主题热度低。论证：两个主题研究内容为污染物处理技术，但研究热度低。

结论二：论文发文时间不同。论证：ZT7 近 10 年都有论文发表，而 HT8 最近 5 年论文被引数量才明显增长，是具有研究潜力的主题。

基于主题具体内容，通过对论文主题热度的分析可以进一步研究主题的价值和关注度，零被引论文的研究主题选择高被引论文中短期的前沿研究或是最近几年才开始关注的话题，由此发现零被引论文善于跟随热点研究。

4.3.2　零被引论文与高被引论文发文期刊对比分析

首先，将同一主题下论文按时间进行排序，然后对所在相同期刊的数量进行统计，再使用统计分析软件 IMB SPSS Statistics 24.0 进行分析。在 SPSS 软件中，以每篇论文为自变量(X)，即横坐标 Paper，以相同期刊上载的论文数量为因变量(Y)，即纵坐标 Quantity，通过散点图初步判断图像走势，运用指数函数对相关数据进行分析。计算结果见图 5。

对相同主题下零被引与高被引论文发文期刊的研究，从以下几个方面进行分析：

从图 5 中的曲线变化看，零被引论文的曲线变化明显，说明零被引论文集中在发文时间较早的期刊上，而高被引论文在相同期刊的数量随着时间的推移变化幅度小，都集中在稳定的数值，其中，HT5 的曲线成正向增长，说明该主题的论文集中在近几年发表的期刊上。

从论文的分布来看，零被引论文分布相对分散，分布在不同的期刊上，期刊种类多；而高被引论文的分布相对集中，有规律，期刊种类少。

从期刊的载文数量分析，零被引的 4 个主题中，期刊载文数量最多的前 5 名期刊是 *Fresenius Environmental Bulletin* (41)、*Environmental Progress & Sustainable Energy* (17)、*Sustainability* (14)、*Journal of Environmental Sciences* (13)、*Frontiers of Environmental Science & Engineering* (13)，其影响因子分别是 0.425、1.672、0、2.937、1.716。这些期刊的专业排名是 Q3、Q4，属于排

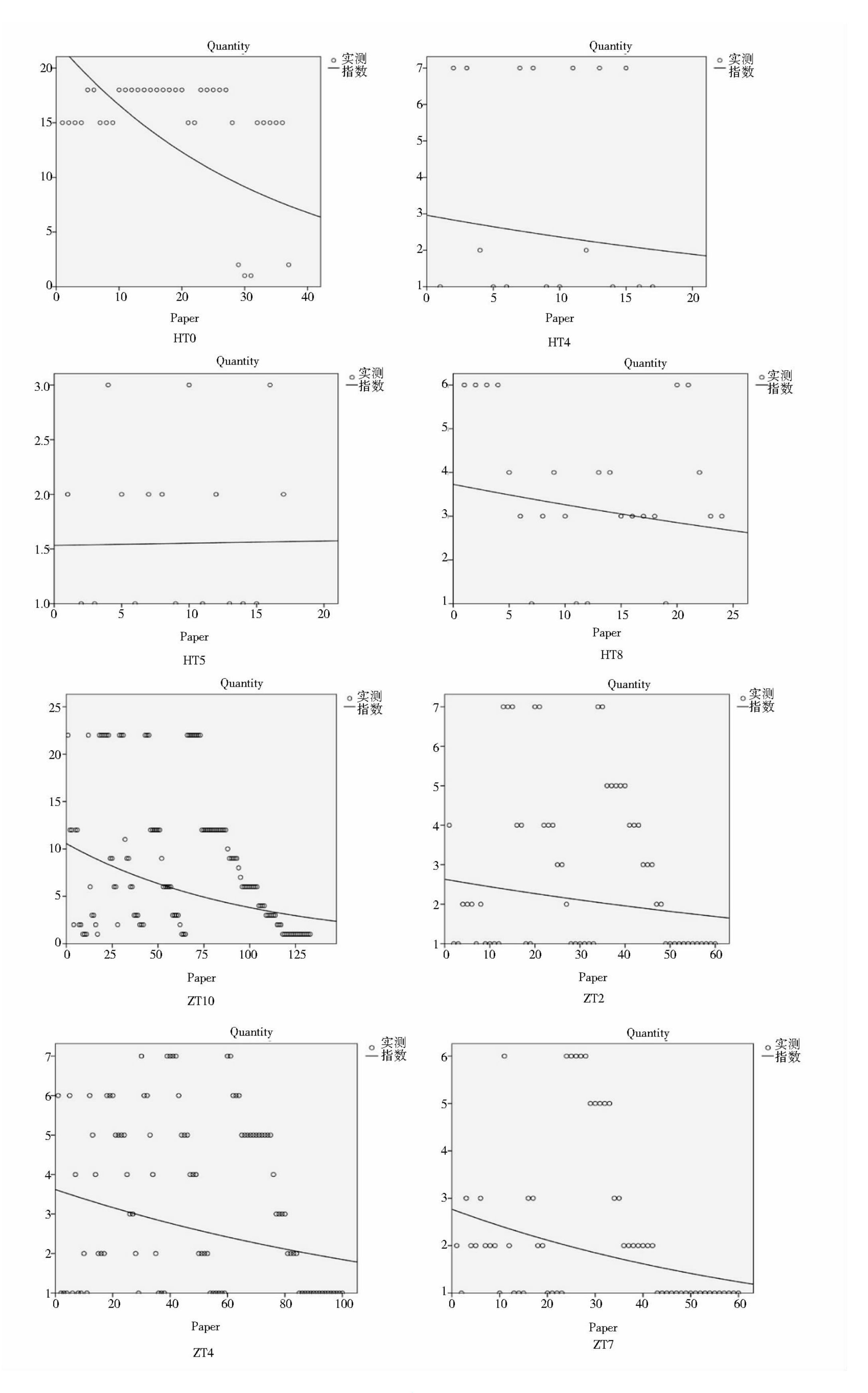


图 5 相同主题下零被引与高被引论文主题的发文期刊变化

名较低、影响力小的期刊。在高被引的4个主题中，载文数量最多的期刊是 *Journal of Hazardous Materials*(33)、*Environmental Science & Technology*(20)、*Science of the Total Environment*(7)、*Water Research*(6)、*Energy & Environmental Science*(4)，其影响因子分别是6.065、3.751、4.900、6.942、5.715、29.518。通过期刊排名表，这些期刊排名为Q1，属于高质量、影响力大的期刊。

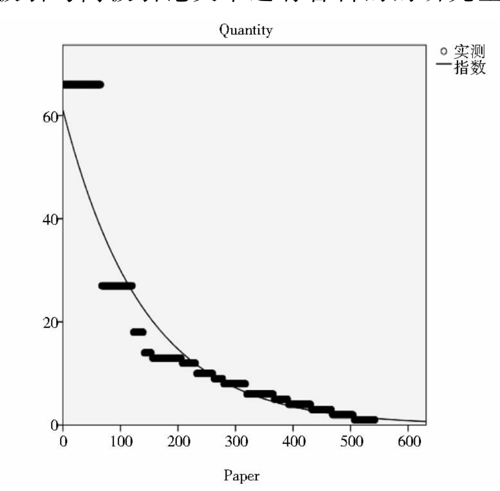
通过4.3.1、4.3.2的特征对比分析，发现零被引论文的研究主题并非是过时的、无用的研究主题，零被引论文的主题紧跟高被引论文的研究热点或者最新出现的主题，以"热门主题"或具有发展潜力的主题为主。从发文期刊分析，零被引论文虽然主题热度高，但发文时间早、发表期刊种类多并且发表在影响因子小的期刊上；高被引论文虽然主题热度低，但发表时间近、发表期刊种类相对集中并且期刊的影响因子高，得到了较高的引用频次。所以在相同主题下，期刊影响因子大小是影响论文不被引用的主要原因。

4.4 零被引与高被引论文不同主题特征对比

在零被引与高被引论文中还有各自的的研究主题，零被引论文是ZT0、ZT1、ZT3、ZT5、ZT6、ZT8、ZT9、ZT11，高被引论文是HT1、HT2、HT3、HT6、HT7、HT9，将从主题热度、发文期刊对零被引论文与高被引论文进行对比分析。

零被引论文中ZT9是高热度主题，ZT0、ZT1、ZT3、ZT6、ZT8、ZT11是一般热度研究主题，ZT5是低热度研究主题；高被引论文中HT1、HT7、HT9是高热度主题，HT3、HT6为一般热度主题，HT2为低热度主题。从主题的研究热度分布来看，高被引论文主要集中在高热度主题，其次是一般热度主题，说明高被引论文的研究主题有很高的关注度，代表了研究前沿；而零被引论文的主题集中在一般热度的研究主题，论文的关注度不高，在论文选题上没有把握当前的研究方向。

不同主题的零被引论文与高被引论文的发文期刊对比，如图6所示，在SPSS软件中，以每篇论文为自变量(X)，即横坐标Paper，以相同期刊上载的论文数量为因变量(Y)，即纵坐标Quantity，主要从以下几个方面进行分析：



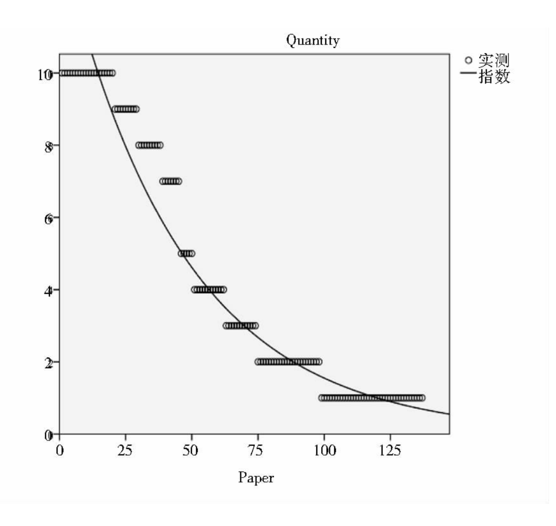


图6 不同主题下零被引与高被引论文主题的发文期刊变化

从图6中的曲线变化看，零被引论文的曲线变化速度快，论文主要集中在尾部的期刊，说明大部分期刊上存在零被引论文但数量少；而高被引论文在相同期刊的数量随着时间的推移变化速度慢，说明高被引论文所在期刊的引用数量差距小。

从论文的分布来看，零被引论文分布相对集中，固定在几种期刊上，期刊种类少；而高被引论文的分布均匀，分布在不同的期刊上。

从期刊的载文数量分析，零被引论文中论文数量最多的期刊是 *Fresenius Environmental Bulletin*(66)、*Environmental Science and Pollution Research*(27)、*Environmental Earth Sciences*(27)、*Polish Journal of Environmental Studies*(18)、*Journal of Coastal Research*(14)，这些期刊的影响因子分别是0.425、2.741、1.569、0.793、0.915。而高被引论文中刊载论文数量最多的期刊是 *Environmental Science & Technology*(23)、*Journal of Hazardous Materials*(22)、*Energy Policy*(14)、*Energy & Environmental Science*(12)、*Science of the Total Environment*(11)，其影响因子分别是6.198、6.065、4.140、29.518、4.9。从期刊影响因子对比分析，高被引论文的期刊影响因子高于零被引论文的期刊影响因子，期刊质量高于零被引论文的发文期刊，零被引论文多发表在影响

因子小的期刊上。

通过以上分析,零被引论文发文期刊相对集中并且发表在影响因子小、知名度低的期刊上,而高被引论文则正好相反,期刊分布均匀并发表在影响因子大、知名度高的期刊上。高被引论文除了发表在影响因子高的期刊外,在研究主题的选择上优于零被引论文,HT1、HT7、HT9 是高热点主题,研究主题新颖,在内容上得到的关注度高;而 ZT0、ZT1、ZT3、ZT5、ZT6、ZT8 等是一般热度主题,研究主题关注度低。因此,在不同主题下,研究主题内容是影响零被引论文存在的主要原因。

5　结语

针对目前研究中主要利用文献计量指标进行零被引论文原因分析,没有深入到文本内容中这一问题,本文利用 PLDA 模型识别高被引与零被引论文摘要中的主题,通过主题相似度计算高被引与零被引论文主题之间的相似度,对比分析零被引与高被引论文在相同主题、不同主题下的主题热度、发文时间、发文期刊的指标变化,进一步揭示零被引论文产生的原因。

在主题相同情况下,发文期刊是影响零被引论文的主要原因。高被引论文以期刊影响因子大、排名较高的期刊为主,高被引论文在一定程度上代表了研究前沿,一些研究人员紧跟其后发表相关主题的论文,并且发表在影响因子不高的期刊上。这种重复已发表在高水平期刊上的“可重复性项目”使得零被引论文数量大增。

在主题不同的情况下,论文主题的选择是导致零被引论文存在的主要原因。高被引论文不仅发表在影响因子高的期刊上,并且研究主题热度高,主题新颖,具有研究价值和指导性作用;而零被引论文大多选择主题热度一般的主题,这些主题具有较好的研究成果和相对成熟的研究体系,缺乏创新性,并且论文集中在影响因子小的期刊上,影响力小。

总之,零被引论文并非是毫无价值的。我们要探究和挖掘零被引论文的价值,不要因为引用次数的限制而忽略了论文本身的价值。由于 Web of Science 数据库对论文收录范围与数量有限,在样本数据获取、统计处理与分析方面难免存在一定的偏差,文中研究结果仅作为一定的参考,部分结论还有待于领域专家验证。

参考文献:

[1] GARFIELD E. Citation indexes for science: a new dimension in documentation through association of ideas[J]. Science, 1955, 122(3159):108 – 111.

[2] GARFIELD E. Citation analysis as a tool in journal evaluation journals can be ranked by frequency and impact of citations for science policy studies[J]. Science, 1972, 178(4060):471 – 479.

[3] 朱梦皎, 武夷山. 零被引现象:文献综述[J]. 情报理论与实践, 2013, 36(8):111 – 116.

[4] 高继平, 潘云涛, 武夷山. 零被引论文的形成因素分析——以光谱学领域零被引论文的国家、机构和主题分布为例[J]. 科技导报, 2015, 33(8):112 – 119.

[5] WEALE A R, MICK B, LEAR P A. The level of non-citation of articles within a journal as a measure of quality: a comparison to the impact factor[J]. BMC medical research methodology, 2004, 4(1):14 – 14.

[6] VAN LEEUWEN T N, MOED H F. Characteristics of journal impact factors: the effects of uncitedness and citation distribution on the understanding of journal impact factors[J]. Scientometrics, 2005, 63(2):357 – 371.

[7] 唐晓莉, 武群芳, 王继民. 论文零被引率与期刊影响力关系的研究——以经济学学科为例[J]. 图书情报工作, 2014, 58(19):100 – 104.

[8] 李美玉, 王硕, 郑德俊. 中文期刊零被引率与期刊关键评价指标相关性分析——以图书情报学科为例[J]. 中国科技期刊研究, 2015, 26(4):399 – 404.

[9] 方红玲. 我国 5 种眼科学中文核心期刊零被引论文特征分析[J]. 中国科技期刊研究, 2014, 25(7):945 – 948.

[10] 魏瑞斌, 王炎. 国外图书情报学期刊的零被引现象研究[J]. 情报杂志, 2015(7):29 – 33.

[11] 杨思洛, 程爱娟. 图情档期刊论文的零被引现象分析[J]. 情报学报, 2015, 34(3):247 – 256.

[12] 胡泽文, 武夷山, 袁军鹏. 零被引研究文献的知识图谱分析——历史发展脉络、主体和高频主题[J]. 情报科学, 2016, 36(3):85 – 91.

[13] 温芳芳. 我国情报学论文零被引的成因及影响因素探析——基于零被引与高被引论文的比较[J]. 情报理论与实践, 2016, 39(4):27 – 31.

[14] 杜新征, 叶文娟, 余茜.《水生生物学报》零被引频次文章特征分析[J]. 编辑学报, 2016(S1):106 – 108.

[15] 赵越, 肖仙桃. 基于主题因素分析的图书情报领域零被引现象研究[J]. 中国科技期刊研究, 2017, 28(7):641 – 646.

[16] 况书梅,伍玉,韩毅. 科研论文零被引的内容因素影响分析[J]. 评价与管理,2017,15(3):36 – 39.

[17] 李贺琼,张小为,傅贤波,等. 10 种外科学综合类期刊零被引论文的特征分析[J]. 中国微创外科杂志,2018,18(4):304 – 308.

[18] BLEI D M, NG A Y, JORDAN M I. Latent Dirichlet allocation[J]. Journal of machine learning research, 2003(3):993 – 1022.

[19] TU Y N, SENG J L. Indices of novelty for emerging topic detection[J]. Information processing & management, 2012, 48(2):303 –

325.

[20] 黄鲁成，唐月强，吴菲菲，等. 基于文献多属性测度的新兴主题识别方法研究[J]. 科学学与科学技术管理，2015(2):34-43.

[21] 刘自强，王效岳，白如江. 多维主题演化分析模型构建与实证研究[J]. 情报理论与实践，2017，40(3):92-98.

[22] 白如江，冷伏海. k-clique社区知识创新演化方法研究[J]. 图书情报工作，2013，57(17):86-94.

[23] Thomson Reuters. Citation thresholds. [2017-06-25]. http://www.sciencewatch.com/about/met/thresholds/.

作者贡献说明：

潘菲：负责数据收集分析和论文撰写；

王效岳：拟定研究命题和思路设计；

白如江：论文框架确定和论文细节修改；

周彦廷：文本数据处理。

An Analysis of Zero-cited and Highly-cited Papers in the Perspective of Research Topics: A Case Study of Environmental Science

Pan Fei　Wang Xiaoyue　Bai Rujiang　Zhou Yanting

Institute of Scientific & Technical Information, Shandong University of Technology, Zibo 255049

Abstract: [**Purpose/significance**] This paper analyzes zero-cited papers in the field of environmental science from the perspective of the subject, to find the differences in the content of articles and external indicators between zero-quoted papers and high-cited papers and reveal the reason for the existence of zero-cited papers. [**Method/process**] Firstly, the PLDA model was used to identify topics that from 260 high-cited papers and 907 zero-cited papers in the domestic environmental sciences database from the Web of Science database. Then the relevance of the topics was found through topic similarity calculation. With the topic popularity used as an internal indicator, the time of publication and the journals used as external evaluation indicators, a comparison analysis of zero-cited papers and high-cited papers was made by combining topical of the papers with external indicators. [**Result/conclusion**] The experimental results show that under the same research topic, the influence of the journal is the main reason that influences the citation of the paper; under different topics, the topic is the main reason leading to zero-cited papers.

Keywords: zero-cited　high-cited　topic recognition　environmental science　comparative analysis　evaluation index